Contents

Pre-flight Probe — Complete Eval v6 (Phase 8 template-lock verdict)

Closes the second open question after v5 ("L43 pre_tool was epiphenomenal — does CoT-Integrity at L55 fare better?"). Phase 8 single-shot bidirectional steering with bf16 amplitude diagnostic delivers a **structural-rigidity null** that converges with Phase 7 on the higher-order claim: **probes detect; mid-layer steering doesn't lever** — but via a distinct mechanism.

After v6, the paper has two complementary epiphenomenal probes from one model, with two diagnosed mechanisms. That's a sharper methodology contribution than v5 alone.

A. Setup

A.1 Probe under test

- Layer: L55 (last attention block before final norm in Qwen3.6-27B)
- Position: last prompt token (post-`<think></think>` injection from `enable_thinking=False` chat template)
- Training data: 240 nb47b captures × `has_think_v1` labels (continuation past auto-injected `<think>`)
- Method: L1-LR C=1.0 on full 5120 dims, balanced class weight
- 4-fold CV AUROC: **0.912** (vs random-feature baseline ~0.50, gap +0.41)

This is a stronger correlative finding than L43 pre_tool. By v5's logic, it'd warrant the same treatment: cheap intervention battery before any causal claim.

A.2 Steering experiment (one-shot, bidirectional)

8 test targets in `ensemble-gated` (RAG context): - 4 **pushup**: baseline `<think>` not emitted, lowest 4 probe scores 0.144–0.171 (probe agrees) - 4 **pushdown**: baseline `<think>` emitted, lowest 4 probe scores within positives 0.698–0.773 (probe agrees but loosely)

Hook: register_forward_hook on `model.model.layers[55]`, fire once on prefill, modify last-position residual: `h[:, -1, :] += $\alpha$ * direction_vec`.

Direction: `clf.coef_ / sc.scale_`, normalized to L2=1. Cast bf16.

Sweep: $\alpha \in$ {-5, -2, 0, +2, +5} on probe direction; $\alpha \in$ {-2, +2} on random direction (control).

A.3 What we observed at $\alpha \in$ {-5..+5}

Across all 8 targets $\times$ 7 (α, direction) combinations = 56 generations: **identical character-by-character output**. Probe and random produce same text as α=0. has_think rate at every condition: 0% (because the `<think></think>` empty pair was already in the prompt — the template forced no-thinking before generation began).

Looked like a hook bug. We ran the diagnostic.

B. The diagnostic — eliminating hook-failure hypothesis

Single target, alpha sweep up to α=+200 (perturbation > residual norm itself):

α (probe)	fired	pre_norm	post_norm	$\Delta\|\cdot\|$	output
0	1	157.73	157.73	0.00	`**The Hunchback of Notre Dame** was released first.\n\n* **The Hunchback of Notr`
+5	1	157.73	157.79	5.00	(identical)
+20	1	157.73	158.91	20.01	(identical)
+50	1	157.73	165.27	50.00	(identical)
+100	1	157.73	186.41	100.00	(identical)
+200	1	157.73	254.25	200.04	(identical)

Random direction at $\alpha \in \{+20, +50, +100\}$: also identical to baseline.

The hook fires. The modification propagates. The residual is genuinely perturbed by Δ=200, +27% above its own norm. Yet the next-token argmax does not flip across 24 generated tokens.

The mechanism is not amplitude. It's structural.

C. Diagnosis — template-locked decision

Last 5 input tokens of every prompt:

```
['\n', '<think>', '\n\n', '</think>', '\n\n']
```

Qwen3 chat template `enable_thinking=False` does not communicate intent through residual content — it **injects the closed `<think></think>` pair as actual prompt tokens** before generation starts. The "no thinking" decision is encoded in the input sequence, not in the model's internal state to be steered.

The L55 last-position residual carries the model's processed encoding of these template tokens. Perturbing it does not retroactively un-emit the `<think></think>` pair already locked into the input. The first generated token must continue from the post-`</think>` cursor — by construction, an answer-text token, not another `<think>`.

In other words: the lever for thinking emission lives in the prompt-construction layer, not in any residual stream layer the probe could be trained on.

D. Convergent verdict — two mechanisms, one regime

Phase 7 and Phase 8 both end at "probe is detection-only, not causal lever". The probes target different signals (capability at tool selection vs CoT emission vs next answer) on different layers (L43 pre_tool vs L55 post-template), and they fail to lever for distinct reasons:

Probe	Phase	AUROC	Failure mode at intervention
L43 pre_tool capability	7	0.83 (gap +0.26)	Δrel $\approx$ 0 across α; probe direction adds **uniform softmax-temperature shift**, not target-specific bias (revealed by control-token normalization)
L55 CoT-emission	8	0.912 (gap +0.41)	Modifying residual at $\alpha = +200$ ($> \|\cdot\|$) produces zero output change; the **decision is template-locked in input tokens**, downstream of where any residual probe could read or write

The unifying claim: **mid-layer linear probes can attain high AUROC by detecting features correlated with the outcome without those features participating in the causal pathway that produces the outcome.** Two regimes confirmed in one model.

This is a cleaner statement than v5 alone, because Phase 8 rules out "amplitude was too weak" — at $\alpha = +200$ with random + probe both null, the residual at L55 last-position simply has no behavioral authority over thinking emission in this prompt regime.

D.5 Phase 8 redux — top-5 concentrated direction confirms structural lock (not dilution)

A first reading of v6 raised the concern that the original Phase 8 null could be a **direction-dilution artifact** rather than structural rigidity: the full-LR steering direction had 5120 components L2-normalized to 1, so per-dim weight was $\sim 1/\$\sqrt{\$5120} \approx 0.014$, with most weight in noise-floor dimensions and only a small fraction on the actual signal-bearing dims.

To rule this out, we recomputed the diff-of-means K-sweep with random-K-matched baselines (`scripts/nb47b_random_k_supp`) and located the paper-grade L55 signal at **K=5 top-diffmeans, AUROC 0.848 vs random K=5 baseline 0.701, gap +0.147**. We then re-ran the amplitude diagnostic on the same target with three directions in parallel:

Direction	active dims	$\alpha = +200$ $\Delta\|\cdot\|$	first 60 generated chars
TOP-5 (paper-grade signal)	5	+199.9	identical to baseline

Direction	active dims	α=+200 $\Delta\|\cdot\|$	first 60 generated chars
RAND-5 (matched control)	5	+200.0	identical to baseline
FULL-LR (Phase 8 original)	5120	+200.0	identical to baseline

12 generations across 3 directions $\times$ {0..200} alpha levels: character-by-character identical. Perturbation post-norm 294 vs pre-norm 158 (+86%) on TOP-5 produced no token change.

This rules out direction-dilution as an explanation. Even when the full $\|\alpha{=}{+}200\|$ is concentrated entirely on the 5 dims that carry the paper-grade signal, the next-token argmax does not flip. The decision is in the input tokens, not in any subspace of the L55 last-position residual.

Paper-grade signal exists at K=5 (probe is real); causal authority does not (lever is upstream in the chat template). The two findings are not in tension — they are the precise statement of "epiphenomenal probe in a template-controlled regime".

Refined interpretation of what the probe reads. The L55 K=5 direction may encode something like *suppressed thinking intent* — a counterfactual signal that says "this prompt would have triggered `<think>` continuation under `enable_thinking=True`, but the auto-injected `</think>` closure has overridden it". The probe distinguishes prompts that retain this latent intent from those that do not. It is detection-only by construction: the lever was already pulled (by the template) before the residual at L55 was computed.

E. Methodology contributions (final list)

The paper contributes three sanity checks for any probe-steering work:

1. **Random-feature baseline + capacity sweep at small N** (Phase 5d $\rightarrow$ eval v2). Catches overparameterization that fakes high AUROC at low N (e.g., Phase 5d K=50 N=17 AUROC=1.000 $\rightarrow$ Phase 6c K=10 AUROC=0.764 once corrected).
2. **Control-token normalization for steering** (Phase 7 $\rightarrow$ eval v5). Catches uniform-softmax-temperature shifts that fake "Δlog-prob(target) shifted by Y nats" headlines. Always report `$\Delta$rel = $\Delta$(target) − mean($\Delta$(controls))`.
3. **Bf16 amplitude + structural-rigidity diagnostic** (Phase 8 $\rightarrow$ eval v6). When a hook produces zero behavioral change, before declaring null, sweep α to multiples of the residual norm. If output is still rigid, the decision lives outside the residual at this layer/position — likely in input tokens for format-like choices in templated models.

All three were one-shot caveats that, before discovery, would have produced confident false-causal claims at minor compute investment.

F. Updated paper claim (post-v6)

Title (final v2): **"Two Forms of Epiphenomenal Probes in Code Agents: Mid-Reasoning Capability and CoT Emission in Qwen3.6-27B"**

Abstract:

> *We train linear probes on residual-stream activations of Qwen3.6-27B during agent rollouts on SWE-bench Pro and probe-gated retrieval on HotpotQA. We obtain two correlative findings: (a) tool-success at L43 pre_tool achieves AUROC 0.83 (gap +0.26 above random-feature baseline) at N=54, and (b) CoT emission at L55 last-prompt-token achieves AUROC 0.91 at N=240. To test causal status, we run*

intervention experiments on each. Both fail, but for distinct reasons. The L43 probe direction adds a uniform softmax-temperature shift, not a target-specific bias — revealed by control-token normalization across `finish` vs `search/execute/write/read/wait`. The L55 probe direction produces zero behavioral change even at α=+200 (>$\|$residual$\|$) because the thinking-emission decision is encoded in the chat template's auto-injected `<think></think>` token pair, downstream of any layer at which the residual could be read. We argue this places linear probes in an explicit epiphenomenal regime when applied to format-like or template-controlled decisions in instruction-tuned models, and contribute three sanity checks (random-feature baseline at small N; control-token normalization for steering; structural-rigidity α-sweep) as standard validations for future probe papers.

This version is more precise than v5: two probes, two failure modes, three methodology checks, all from one model and ~$8 of intervention compute.

G. Updated SDK plan (final v0.1, no changes from v5)

`agent_probe_guard_sdk_plan.md` Section 3.1 stays as-is: - L43 pre_tool, K=10 features, top-10 by diff-of-means - skip / escalate / proceed gating only (no boost mode) - README copy: "reads the model's mid-reasoning capability assessment; does not modify behavior"

Phase 8 confirms the same posture for any probe-gated CoT control ideas: detection-only is the right framing.

H. Updated decision tree

Outcome at Phase 6 N=99	Probability (v6)	Paper	SDK
A: AUROC $\geq$ 0.80, gap $\geq$ +0.20	55%	accept (workshop, two-probe story)	ship v0.1 detect-only
B: AUROC 0.70-0.80, gap +0.10 to +0.20	30%	accept w/ revisions	ship v0.1 conservative threshold
C: AUROC 0.55-0.70, gap +0.05 to +0.10	10%	honest-negative + methodology-only	no SDK
D: AUROC < 0.55	5%	methodology-only	no SDK

Outcome A bumps from 50% to 55% — the L55 N=240 finding (AUROC 0.91) gives the paper a stronger second probe regardless of N=99 outcome at L43. Even if L43 collapses at scale, the L55 finding plus the two-mechanism diagnosis carries the paper.

I. What remains

1. Phase 6 N=99 trace collection completes (~4-5h remaining) — automatic
2. Re-run methodology sweep at N=99 — confirm L43 gap holds
3. Permutation null at K=10 (1000 label shuffles) — proper p-value for L43
4. Bootstrap CI 1000 iter at K=10 — replace current 200-iter for L43
5. Permutation null at full-feat L1 (1000 shuffles) — proper p-value for L55
6. **(Optional, future work)** SAE-decoded steering on L55: pick SAE features in qwen36-27b-sae-papergrade with high cosine to L55 probe direction, steer those instead. Yap (2026) showed SAE features can lever where linear directions cannot. ~$5, 4-6h.

7. **(Optional, future work)** Token-level intervention: replace `<think>\n\n</think>\n\n` injection with a steerable single-token slot, retrain probe at that position, repeat experiment. Tests whether L55 probe can lever when the decision is genuinely in the residual.
8. Paper draft consolidation — eval v3 + v4 + v5 + v6 → final paper sections.
9. SDK v0.1 build (~5 days post-submit).

Items 6 and 7 are deferred future work in the paper, not blockers.

J. Cost ledger

Phase	Compute	Outcome
Phase 5d (over-param N=17)	~$3	Caught by random-feature check → eval v2
Phase 6c (methodology sweep)	~$2	L43 pre_tool K=10 real signal → eval v4
Phase 7 (3-source causality)	~$2	L43 epiphenomenal → eval v5
Phase 8 (CoT bidirectional + diagnostic)	~$3	L55 epiphenomenal (structural) → eval v6
Phase 8 redux (random-K + top-5 retest)	~$1	TRUE L55 signal at K=5 (gap +0.147); structural lock not dilution → v6 §D.5
Total intervention budget	**~$11**	Two probes, two mechanisms, three methodology checks, K=5 paper-grade signal

For reference: a single Docker SWE-bench Pro eval with 100 traces costs ~$15. The full intervention battery cost less than running the eval once and produced the science that justifies (or doesn't) running it.

K. The v6 lesson distilled

For any linear-probe paper on instruction-tuned models with chat templates:

- **Detection ≠ causation** (v5 lesson holds).
- **For format-like or template-controlled outcomes, the decision is often in input tokens, not residuals.** Mid-layer steering at any α cannot reach it. The structural-rigidity diagnostic catches this in <60s of compute.
- **Two failure modes for "probe doesn't lever"** are now documented in this work: softmax-temperature artifacts (Phase 7) and template-locked decisions (Phase 8). Future probe-causality work should report against this taxonomy.
- **Detection ≠ causation is an instance of eval-distribution overfitting.** Anthropic Alignment (2026, *Teaching Claude Why*, https://alignment.anthropic.com/2026/teaching-claude-why/) document the same pattern at the alignment-training level: in-distribution metrics can pass while held-out automated auditing fails. Linear probe AUROC is the in-distribution metric here; behavioral steering at $\alpha \gg \|h\|$ with both probe and random direction is the held-out audit. The two-mechanism taxonomy above is the failure-mode map for that audit in mech-interp tooling.
- **OpenAI Alignment (2026, *Accidental CoT Grading*, https://alignment.openai.com/accidental-cot-grading/) report the behavioral counterpart**: when CoT text was inadvertently included in RL reward, "surface-level CoT properties were steerable under sufficient pressure, but more specific monitor-relevant shaping was harder to induce". This matches our two-mechanism taxonomy: surface softmax-temperature class is steerable (Phase 7); deeper template-locked or input-token-controlled decisions

are not (Phase 8). Three labs (Anthropic, OpenAI, OpenInterp) now converge on the same structural finding from three substrates (alignment training, RL CoT reward, residual-stream probes).

The cumulative lesson across v2 $\rightarrow$ v6: every cheap sanity check we considered skipping turned out to flip a confident finding into an honest one. The discipline was net-positive at every step.